Journal of Physics: Conference Series

PAPER • OPEN ACCESS

An analysis of Goodhart's law toward a shared conceptual framework of measurement across the sciences

To cite this article: L Mari and R Ruffini 2018 *J. Phys.: Conf. Ser.* **1065** 072022

IOP Conf. Series: Journal of Physics: Conf. Series **1065** (2018) 072022 doi:10.1088/1742-6596/1065/7/072022

An analysis of Goodhart's law toward a shared conceptual framework of measurement across the sciences

L Mari[1], **R Ruffini**[2]

[1] School of Industrial Engineering, Università Cattaneo, Castellanza, VA, Italy

[2] Università Statale, Milano, Italy

E-mail: lmari@liuc.it

Abstract. An analysis of Goodhart's law and an exploration of the conditions of its validity reveal significant analogies and differences between the measurement of physical and non-physical properties, where the so-called Hawthorne effect can be interpreted as a generalized loading effect. It is an opportunity to test the feasibility and the usefulness of a shared conceptual and lexical framework of measurement across the sciences.

1. Introduction

The study of the analogies and the differences between the measurement of physical and non-physical (psychological, social, etc) quantities, and more generally properties, is a starting point toward the development of *a shared conceptual framework of measurement across the sciences*. It is a worthwhile target in our growingly complex society, in which socio-technical systems are widespread and more and more problems of both physical and social (a positive, short term for "non-physical") measurement need to be solved. While sufficiently significant analogies prove the feasibility of such development, the evidence of diversities is not less important, for the information it conveys that distinct sciences adopt different solutions to the same problems or that they differently emphasize some aspects of the complex process of measurement.

Due to its long and well established tradition (and perhaps as a consequence of what has been called "physics envy", leading researchers to mimic methods drawn from physical sciences even when they are not appropriate to the context [1]), physical measurement would plausibly play the leading role in most stages of this development. On the other hand, social measurement has specificities which, if properly taken into account, may enrich the framework, making it more general and encompassing, a precious feature in a situation in which socio-technical complexity is more and more expected to be dealt with by acquiring information on both physical and social properties of the objects under consideration (we use the term "object" in a broad sense, to designate also events, processes, individuals, etc: basically, an object is any entity bearing properties). In this perspective, a peculiar

[1] To whom any correspondence should be addressed.

aspect of social measurement, sometimes just unknown in physical measurement, is related to the so-called *Goodhart's law* (GL for short henceforth) originally stated as [2]

GL1: *any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes*

and commonly presented as [3]

GL2: *when a measure becomes a target, it ceases to be a good measure*

An analysis of this law and an exploration of the conditions of its validity are at the same time an exercise of the possibility of a mutual understanding of physical and social measurement scientists and an opportunity to test the feasibility and the usefulness of a shared conceptual and lexical framework of measurement across the sciences.

2. Preliminary analysis

The logical structure of GL is

if *premise* then *conclusion*

where *a measure becomes a target* is the condition, stated as sufficient, for the truth of the consequence, *a measure ceases to be a good measure*. Interestingly for our purposes, formulations GL1 and GL2 do not include any domain-related specification: is then GL claimed to be true independently of the context, hence for both physical and social measurement? or should some further conditions, that remained implicit in these formulations, be elicited?

Preliminarily, the ambiguity should be removed about the actual meaning of the term "measure" in GL2 (the *International Vocabulary of Metrology* (VIM) [4] avoids using the term "measure" because of its ambiguity). In principle, "measure" could refer to the process of measurement, and therefore GL would mean that having the purpose of measuring is sufficient to worsen the process. Given that measurement is in fact a designed-on-purpose process, this interpretation would make the law systematically false. Moreover, GL1 shows that "measure" refers here to a quantity and the concept <to become a target> should be made explicit. Let us then propose a slightly refined formulation:

GL3: *if a quantity becomes a target for control purposes then it ceases to be a good quantity to be measured*

While GL has a normative intent (i.e., don't choose a quantity as a target for control purposes if you want to maintain it a good quantity to be measured), its structure shows its descriptive nature, such that the law is in principle either true or false. In order to identify the (necessary or sufficient) conditions of truth of GL, we need to specify what characterizes a good measurement.

According to the VIM, a measurement is a "process of experimentally obtaining one or more quantity values that can reasonably be attributed to a quantity" [4], where then by means of measurement an information entity (i.e., one or more quantity values) is "reasonably attributed" to an empirical entity (i.e., the quantity of an object). Measurement is a process of acquisition, formalization, and presentation of information about an empirical quantity (and more generally an empirical property) of an object, and a good measurement is then expected to produce information describing with sufficient reliability the state of the object under measurement at the moment of the measurement relatively to the measured quantity. Weak interpretations of this conditions are widespread, emphasizing the representational nature of such a description: if, e.g., the length L_a of the object a has been measured to be 1.2345 m, it is because the value 1.2345 m represents the empirical quantity L_a. In this view, a good measurement may be supposed to be a process whose results satisfy the representational condition [5] such that, e.g., if the object a is perceived to be warmer than the object b then the value of temperature

attributed to *a* must the greater than the value of temperature attributed to *b*. We rather endorse a strong interpretation, that a good measurement produces a measured value corresponding to a quantity which is empirically indistinguishable from the measurand. In this view the basic relation *measurand* = *measured value*, e.g., L_a = 1.2345 m (not considering measurement uncertainty here) is an equation, not just a representation. If the length of *a* has not changed in the meantime, this is the result of a good measurement if any object whose length is 1.2345 m is empirically indistinguishable from *a* by their lengths. This strong, descriptive interpretation entails the weak, representational interpretation, but not vice versa: a good description also provides a good representation, but not vice versa. Hence, any conclusion that we reach also applies in the representational case.

In this sense multiple reasons are possible for a measurement not to be a good measurement. In particular, for the measurement result to be exploited as a decision making tool a target uncertainty should be decided, and a measurement is not good if it produces values with a standard uncertainty greater than such target uncertainty (we are adopting here the terminology of the VIM, as framed in the *Guide to the expression of uncertainty in measurement* (GUM) [6]). Precisely in the context of decision making, let us consider the case of a sensor in a closed-loop feedback control system, such as the temperature sensor in a thermostat controlling a heater, where the measurement uncertainty of the thermometer should be less than the target uncertainty, chosen in function of the minimum acceptable resolution of the controller. Here GL *premise* = *a quantity becomes a target for control purposes* clearly applies, the general quantity being temperature, of which the set point specified through the user interface of the thermostat is an instance. Since GL *conclusion* = *temperature ceases to be a good quantity to be measured* is plainly false in this case, we must conclude that GL is once again false. In other terms, it is false that in general values of a quantity lose their descriptive quality as soon as that quantity becomes a target for control purposes, as instead claimed by GL.

In order to recover the validity of GL we need some further analysis, aimed at specifying some conditions to be added to the premise:

if *premise* and *conditions* then *conclusion*

3. When Goodhart's law is valid

The fundamental, structural reason of potential reduction, and in the extreme case loss, of descriptive quality of a measurement is already acknowledged in the tradition of physical measurement: it is about the situations in which *the very act of measuring affects the state of the object under measurement*. For a measurement to be performed the measuring instrument (MI) must be such that its state can change in response to the interaction with the object under measurement (OuM): the point is whether and when there is also a reverse interaction, from the MI to the OuM. While sometimes the OuM state can be considered to be practically unaffected by the interaction with the MI (e.g., the state of a star surely does not change because of a radiotelescope detecting the electromagnetic radiation emitted by the star), in some cases the coupling of the OuM with the MI cannot be neglected. In order to characterize such cases, let us exploit the traditional distinction, random vs systematic, but apply it to model whether the OuM state is randomly or systematically affected by the interaction with the MI, thus focusing on the latter. In electrical measurement this is known as *loading effect* (e.g., any voltmeter whose internal resistance is not sufficiently high reduces the voltage across the resistor under consideration), systematically biasing the measurand which as a consequence "ceases to be a good quantity to be measured for control purposes" (such that, e.g., the heater would maintain a temperature systematically higher or lower than the set point). On the other hand, if a model of the MI and its

interaction with the OuM is available, then the loading effect can be corrected by appropriately calibrating the MI. *This assumes the knowledge not only of the MI behavior but also, and critically, of the way the state of the OuM is affected by its interaction with the MI*. This might be unproblematic in physical measurement (the equivalent resistance of the parallel circuit of the resistor under measurement and the voltmeter is easily computed), but it is a key issue in social measurement, where a loading effect could be identified but implausibly modeled (e.g., employees might change their behavior when they are aware of their being evaluated / measured according to it, but the amount of the change usually depends on so many subject-related factors that it can be hardly predicted).

Hence, if the OuM (i) is informed / aware of its/her/his condition of being under measurement, and (ii) can change its/her/his state as a consequence, and (iii) has an interest to do it, then the knowledge of measurement might induce the OuM to systematically change its/her/his state to adapt to given / supposed expectations. Such a set of conditions is sometimes synthesized as the "Hawthorne effect", "a type of reactivity in which individuals modify an aspect of their behavior in response to their awareness of being observed" [7], thus a sort of *generalized loading effect* (a delicate example is provided by "the audit culture of universities – their love affair with metrics, impact factors, citation statistics and rankings – [that] does not just incentivize [a] new form of bad behaviour. It enables it." [8]; for a metrological analysis of this subject see [9]). This leads us to a better formulation of GL:

> GL4: *if a quantity becomes a target for control purposes and the object under measurement is affected by the generalized loading (i.e., Hawthorne) effect then it ceases to be a good quantity to be measured*

In the encompassing perspective of measurement as a core component of decision making, taking Goodhart's law into account generates then a new dimension to a possible conceptual framework of measurement across the sciences: measurement as a tool for inducing a state transition in the object under measurement.

4. References

[1] von Hayek F 1989 The pretence of knowledge *The American Economic Review* **79** 3–7
[2] Goodhart C 1981 Problems of monetary management: the U.K. experience *Inflation, depression, and economic policy in the West* ed A S Courakis (Lanham, MD: Rowman & Littlefield) pp 111–146
[3] Wikipedia *Goodhart's law* en.wikipedia.org/wiki/Goodhart's_law
[4] JCGM 200:2012 *International Vocabulary of Metrology – Basic and general concepts and associated terms* (VIM) 3rd ed (2008 version with minor corrections) Joint Committee for Guides in Metrology (www.bipm.org/en/publications/guides/vim.html)
[5] Krantz D H, Luce R D, Suppes P, Tversky A 1971/1989/1990 *Foundations of measurement* vols. 1–3 (New York: Academic Press)
[6] JCGM 100:2008, *Evaluation of measurement data – Guide to the expression of uncertainty in measurement* (GUM) Joint Committee for Guides in Metrology (www.bipm.org/en/publications/guides/gum.html)
[7] Wikipedia *Hawthorne effect* en.wikipedia.org/wiki/Hawthorne_effect
[8] Biagioli M 2016 Watch out for cheats in citation game *Nature* **535** 201
[9] Petri D, Mari L, Carbone P 2015 A structured methodology for measurement development *IEEE Trans Instr Meas* **64** 2367–2379